PB86-197829

Computer Assisted Studies of Structure-Activity Relations Using Pattern Recognition

(U.S.) National Science Foundation
Washington, DC

Prepared for

Health Effects Research Lab.
Research Triangle Park, NC

1983

U.S. Department of Commerce
National Technical Information Service
NTIS

149-10

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before comp)

1. REPORT NO.	2.	PB86-197829
EPA/600/J-83/340		

4. TITLE AND SUBTITLE
COMPUTER ASSISTED STUDIES OF STRUCTURE-ACTIVITY RELATIONS USING PATTERN RECOGNITION (Journal Version)

5. REPORT DATE
1983

6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)
Peter C. Jurs

8. PERFORMING ORGANIZATION REPORT NO.

9. PERFORMING ORGANIZATION NAME AND ADDRESS
National Science Foundation, Washington, DC 20550

10. PROGRAM ELEMENT NO.

11. CONTRACT/GRANT NO.
CR811037

12. SPONSORING AGENCY NAME AND ADDRESS
Health Effects Research Laboratory - RTP, NC
Office of Research and Development
US Environmental Protection Agency
Research Triangle Park, NC 27711

13. TYPE OF REPORT AND PERIOD COVERED

14. SPONSORING AGENCY CODE
EPA/600/11

15. SUPPLEMENTARY NOTES
REF: Drug Information Journal, 17(4):219-229, November 1983

16. ABSTRACT

Structure-activity relationships (SAR) can be investigated for large sets of biologically active compounds using computer-assisted methods. The combination of chemical structure information handling, molecular mechanics model building, substructure searching, and related methods along with pattern recognition and statistical analysis provided an approach to SAR studies. An interactive computer software system, called ADAPT, has been implemented to automate this approach to SAR studies and to allow its application to large sets of organic compounds. A number of SAR studies have been performed using ADAPT, including studies of pharmaceuticals, anti-tumor compounds, olfactory stimulants, and genotoxic compounds (mutagenic or carcinogenic compounds). Results of two studies are discussed in detail: a study of 9-anilinoacridine anti-tumor compounds, and a study of N-nitroso carcinogens.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field.Group

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS *(This Report)*	21. NO. OF PAGES
RELEASE TO PUBLIC	UNCLASSIFIED	12
	20. SECURITY CLASS *(This page)*	22. PRICE
	UNCLASSIFIED	

EPA Form 2220-1 (Rev. 4-77)

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

EPA/600/J-83/340
JOURNAL ARTICLE

Drug Information Journal, Vol. 17, pp. 219-229, 1983
Printed in the USA. All rights reserved.

0092-8615/83 $3.00 + .00
1984 Drug Information Association Inc.

COMPUTER ASSISTED STUDIES OF STRUCTURE-ACTIVITY RELATIONS USING PATTERN RECOGNITION

PETER C. JURS
National Science Foundation
Washington, DC

Structure-activity relationships (SAR) can be investigated for large sets of biologically active compounds using computer-assisted methods. The combination of chemical structure information handling, molecular mechanics model building, substructure searching, and related methods along with pattern recognition and statistical analysis provided an approach to SAR studies. An interactive computer software system, called ADAPT, has been implemented to automate this approach to SAR studies and to allow its application to large sets of organic compounds. A number of SAR studies have been performed using ADAPT, including studies of pharmaceuticals, anti-tumor compounds, olfactory stimulants, and genotoxic compounds (mutagenic or carcinogenic compounds). Results of two studies are discussed in detail: a study of 9-anilinoacridine anti-tumor compounds, and a study of N-nitroso carcinogens.

Key Words: Structure activity; Pattern recognition; Carcinogen; Anti-tumor

I AM GOING to describe briefly the methodology we use to perform structure-activity relationship (SAR) studies. The approach is implemented in the ADAPT computer system, which is a set of software tools that can be used for performing a variety of structure-activity studies. I will be emphasizing a different aspect to SAR than the other talks in this symposium because there are some differences between the class of problems that can be attacked using the methodology we use and the class of problems that can be attacked by some of the alternative methods.

Application areas for SAR include pharmaceutical design, agricultural chemicals, studies of chemical and genetic toxicity, and also other physical or biological properties. We are within the constraints of a relatively small academic research group with a limited amount of manpower; we try to study topics from diverse areas with a diverse set of collaborators.

Figure 1 shows our approach to SAR studies. This is an empirical approach to structure-activity studies. The reason for using an empirical approach is as follows. When the fundamental, molecular level mechanisms of an interaction with a receptor are known, then they can be exploited. And when the fundamental mechanisms are unknown, you have many options on methodologies to use.

In this particular instance, I am starting from the presumption that although molecular structure and biological activity are linked causally, the causal link is not under-

Presented at the Drug Information Association Workshop on Computer Assisted Chemistry in Drug Design, February 1983, Philadelphia.

Reprint address: Peter C. Jurs, National Science Foundation, Washington, DC 20550.

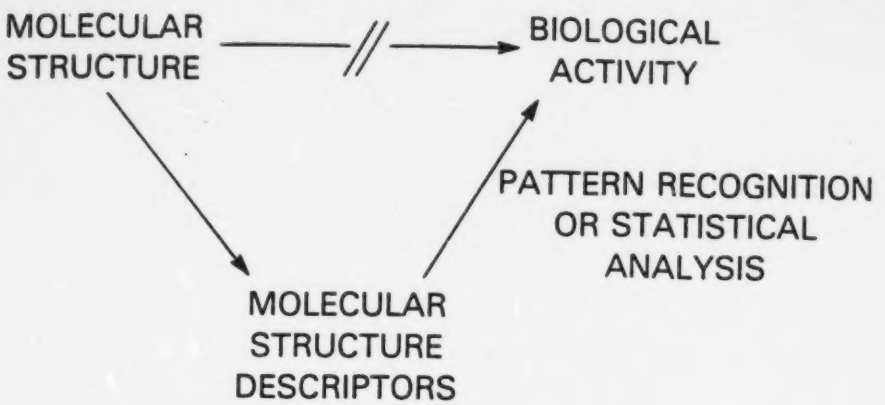

FIGURE 1. Structure-activity studies using pattern recognition.

stood. The broken arrow represents this. Thus, we have to use correlative methods in which molecular structures are entered and stored in a computer system and each structure is represented by calculated molecular structure descriptors. The resulting matrix of data—with one side labeled by compound number and the other side by descriptor number—is analyzed using either pattern recognition or statistical analysis, as the data set characteristics and as the goals of the project demand.

A variety of structure-activity methods exists. The quantitative approach, typically using regression analysis, has come to be known as Hansch analysis after Corwin Hansch. Here, the level of activity of the compounds being studied is known and can be predicted with a quantitative model. Another of the approaches to SAR—structural-mechanistic SAR—involves studies of a fairly small number of compounds fairly intensively. Conformational analysis in its various manifestations, surface representations perhaps using color graphics, and other graphically-oriented methods constitute a set of techniques that can be used to study a small number of compounds intensively. We will focus here on qualitative SAR, where one will typically have only the class of activity of the compounds: active/inactive or highly/moderately/slightly active. For large sets of compounds, carcinogenic or mutagenic potential are examples of class or category data that are often available. Here one must use pattern recognition and statistical analysis to try to develop discriminants separating categories of data rather than trying to predict activities quantitatively.

The premises involved in this approach to SAR are as follows. The first premise is that molecular structure and biological activity are related. Second, we presume that one can represent structures of compounds by sets of molecular descriptors, suitably chosen. What constitutes an adequate description is not known in advance; if it were, then you would not be doing a correlative study.

Given the second premise, then the third is that one can find a relationship that connects structure and activity. Either pattern recognition techniques or statistical techniques are applied to a set of compounds that have gone through some kind of standardized bioassay or physical property measurement. Typically, we are seeking a discriminant that will separate the activity classes from each other. The method relies on having large sets of compounds that have been studied in standardized bioassays or physical property measurements.

Given the relation between structure and

activity, then one can extrapolate it to untested compounds, that is, predict. This could be used to guide a synthesis project in cooperation with other methods of decision-making. It could also be used to allocate resources in a testing program.

The approach just described has been implemented in the computer system we call ADAPT (for Automated Data Analysis using Pattern recognition Techniques).[1] A flow chart for ADAPT is shown in Figure 2. It is a set of software tools, written in FORTRAN, that runs on a minicomputer, and that supports the four basic sets of activities. Because we are in an academic environment, we do not have access to large company files from which to draw problems, so we enter and store our molecular structures individually by sketching them on a screen of an interactive terminal. We have in the system a three-dimensional molecular modelling system for cases in which it makes sense to build molecular models. It is a standard molecular-mechanics model builder. Strain energy is defined in the usual way with respect to intramolecular interactions, which are functions of the atomic coordinates, and then a pattern searching routine is used to minimize this highly nonlinear strain energy function. Reasonably high quality three-dimensional molecular models are generated quite quickly. We display models graphically for visual checking, using a number of algorithms drawn from the literature or implemented by us. Not all of our studies use three-dimensional models of the compounds because of the potential flexibility of some classes of compounds.

The heart of the system is the generation of molecular structure descriptors. If good descriptors can be generated in the central phase of a problem, then the pattern recognition or statistical phase is fairly easy. If good descriptors cannot be developed, the analysis phase is impossible. This is where the creativity and the difficult part of these SAR projects fall.

The goal, of course, is to represent the important features of the structures. What constitutes an important feature? We don't know in advance. The way to determine what is important is to determine whether the set of descriptors being analyzed is sufficient to support a powerful discriminant. We only have indirect information regarding importance, namely, a list of compounds, their structures, and their observed activities. We also have our insights from the literature regarding metabolism, suspected mode of action, etc. So only indirect information is available to guide us in this search for important descriptors. To compound the problem, there are a vast number of possible descriptors, so one has to watch out for statistical traps. Without

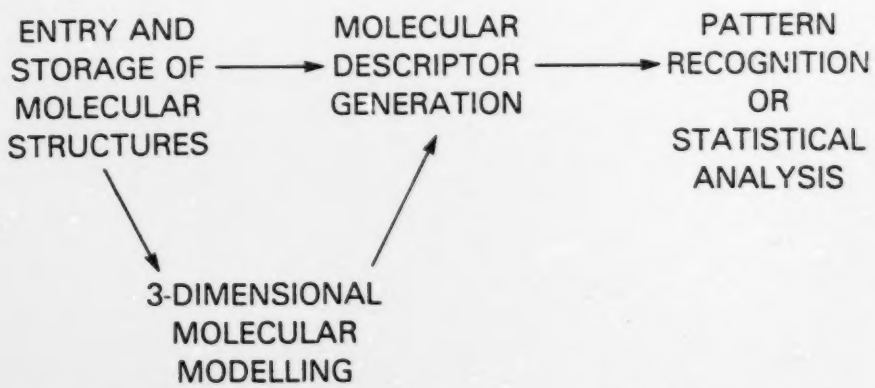
ENTRY AND STORAGE OF MOLECULAR STRUCTURES
MOLECULAR DESCRIPTOR GENERATION
PATTERN RECOGNITION OR STATISTICAL ANALYSIS
3-DIMENSIONAL MOLECULAR MODELLING

FIGURE 2. Overview of the ADAPT computer software system

going into detail here, suffice it to say that we are aware of these statistical traps and how to avoid them. So this descriptor development phase is essentially a guided search.

We use a pragmatic approach in searching for descriptors, and we tend to be willing to use what will work for us. A more scientific way of putting this is that we use ideas garnered from a number of different relevant fields in the generation of descriptors. A list of the fields that have been brought to bear on the generation of descriptors by us and by others who calculate molecular structure descriptors includes mathematics, toxicology, physical organic chemistry, pathology, medicinal and pharmaceutical chemistry, chemical engineering, computer science, molecular pharmacology, and quantum chemistry. An alternative way of looking at the situation is by considering some of the various descriptions of events that combine to produce biological activity. They include chemical reactivity, enzyme-substrate interactions, binding, solubility, transport, electronic properties, and so on. The problem is to generate a set of descriptors that simultaneously represent a spectrum of properties of the molecules of interest, and are as a group sufficient to support the activity differentiation into categories.

Now, in practice, we generate descriptors with the ADAPT system using procedures that we have developed through experience and by applying chemical reasoning to the problem at hand. The exact sequence of descriptor development is utterly dependent on the problem being attacked. The kinds of descriptors that ADAPT supports fall into four general classes:

1. Physico-chemical descriptors, of which log P and molar refractivity would be two examples.
2. Topological descriptors, depending only upon the topology of molecules, which include substructure counts, molecular connectivity, and substructure environments. The substructure count descriptors code only how many times a substructure is present in the molecule being coded. But the way in which they are bonded into the structures of interest—that is, their immediate environments—can also be important. The substructure environment descriptors are an attempt to capture this type of information.
3. Geometrical descriptors describe molecular shape, volume, and surface area; they are steric descriptors.
4. Electronic descriptors. We have implemented some σ charge and extended Huckel calculations for generating relatively simple electronic properties.

Once again, the hallmark of this approach is that the activity is considered to be related to a spectrum of properties of the molecule rather than any single class of descriptor, let alone any single individual descriptor.

In working on a particular SAR problem, we typically go through a sequence involving two very unequal steps. The first step involves calculating once and only once a fairly large number of topological descriptors and geometrical descriptors which are fixed, given the data set. These include fragment descriptors, molecular connectivity, path descriptors, molar refractivity, log P, shape parameters, volume, and surface area. We calculate these descriptors and then use standard statistical methods to prune the resulting set to a minimum set. The objective here is to prune to a minimum set of descriptors that contain the maximum amount of information possible—conflicting goals. This set of descriptors forms the basic description of the data set under study. Then in the second, iterative part of an investigation, one looks at the molecules in the data set and focuses on substructures that are present. We calculate either substructure count descriptors, counting the occurrences of substructors; or substructure environment descriptors to look at how substructures that characterize these

molecules of the data set are imbedded within the structures.

One of the descriptors that can be calculated is log *P*. We have constructed several versions of a program to estimate log *P* based on the constructionist approach developed by Leo, Hansch, and co-workers at Pomona College.[2] It is a weighted, linear summation of fragment constants with proper correction factors. It implements a set of ad hoc rules developed by carefully measuring log *P* values for small compounds and making the assumption that the fragment constants could be extrapolated to larger compounds. For molecules without too much functionality, this estimation procedure works rather well.

Given a set of compounds entered and stored and possibly modelled, given the ability to generate descriptors for them, then the last stage of the SAR study is the analysis stage. This is where statistics and pattern recognition come in. Once each molecule in a data set has been described by a set of relevant descriptors, then studying that matrix of data by pattern recognition or statistical analysis forms the final quarter of the system. The last quarter of the system can, in fact, stand alone, and is being used currently in my laboratory for a number of analytical chemistry studies, such as fingerprint analysis of complex gas chromatographic traces of biological materials.

Fundamentally, we are representing compounds as points in a space. As an example, consider a set of compounds of known activities where each compound was represented by just three descriptors. Then one could plot a particular compound as a point in three-dimensional space, and a different compound in the data set with the same three descriptors, but different values for them, would be represented by a different point in the same space. And if there were 100 compounds in a data set, and each was represented by the same three descriptors, they would be represented by a swarm of 100 points in this three-dimensional space. If one was studying a complicated biological activity such as carcinogenic potential, the chances are that the points representing the carcinogenic and the benign compounds would be overlapped in the swarm of points with no differentiation, because there are no three descriptors that could separate them. But there is no reason to restrict the representation to three descriptors per compound; we can have many more. Of course, they cannot be plotted any more on a cathode ray tube (CRT) screen. We can represent the compounds by *n* descriptors each and thereby devise an *n*-dimensional space.

Pattern recognition methods are nothing but a set of computer-based methods that allow one to investigate clustering of data represented in this formalism. The source of the data here is SAR, or structures represented by molecular descriptors, but the pattern recognition methods are mathematical formalisms, and they are unrelated to the source of the data. Pattern recognition is, in fact, a complete field of study unto itself.

We can represent structures as points in space and investigate them with pattern recognition methods, and ADAPT has within it perhaps a dozen or 15 different pattern recognition methods drawn from all the major classes of pattern recognition methods which have been reported in the literature. One class of pattern recognition methods involves display, where one attempts to map the points in multidimensional space into two-space for direct human viewing. Another class of pattern recognition methods is discriminant development, where the objective is to separate classes of compounds one from another. In terms of the mathematics, that means finding subregions of the space where points common to the carcinogenic compounds cluster separately from the points common to the benign compounds. Discriminants can be developed based either on parametric or nonparametric methods. There are also clustering methods and modeling methods available.

The studies we work on most often are

category type problems. We usually use nonparametric discriminant development methods because the statistics of the data that we study are not well characterized by quantities such as means and covariance matrices.

Typically, we are seeking discriminants that can slice through a high-dimensional space separating clusters of carcinogenic compounds from benign compounds. If a discriminant can be found that completely separates the two classes, then the two classes are said to be linearly separable. This is a desirable case. If in a real study one can demonstrate this kind of linear separability, then it has been shown, after the fact, that one has an adequate set of descriptors. It cannot be predicted in advance what an adequate set is, but it can be shown after the fact.

Very often, in real studies, there are overlaps between the point sets; there is no place to put a simple surface through the center that generates complete separation. In such a case, the objective would be to find better descriptors by some means. In the absence of finding better descriptors, then one generates a discriminant going through the central region, which makes as few classification errors as possible. In other words, use a pragmatic approach when there is no alternative.

Given the case with complete separability, or given a case where separability is not complete but one has put a discriminant through the central region anyway, that constitutes having found a relation based on a set of adequate descriptors which separates. Therefore, one can use that discriminant to predict the classes of new compounds. This is because a new compound would be represented by the same set of descriptors and would be a new point put in the space; it would be classified with respect to the discriminant as being in one class or the other. I have over-simplified quite a bit here to be brief, but these are the fundamentals of the pattern recognition approach.

In the case where quantitative biological or physical property data are available, one can use multiple linear regression analysis to form quantitative models. ADAPT has a statistical package including a multiple linear regression analysis program. In the case where we have studies involving quantitative data and where we wish to generate a quantitative model, that can be done too. ADAPT is an integrated system that allows the user to do a lot of different things depending on the goals of the project. The examples I show will focus on generation of discriminants, but that is not the only kind of study that can be done.

To reiterate: Application areas for molecular structure–biological activity studies have been the focus of what I have said so far, but these methods are also applicable to the case of looking for a molecular structure–physical property relationship. But in the biological activity area we can study pharmaceuticals, agricultural chemicals, chemical communicants—that is, the odor of organic chemicals—and genetic and chemical toxicity.

I want to give several case histories to show briefly the kinds of studies to which such a system can be put. I have two studies to discuss. Taken together they do not illustrate the full breadth of applicability of this methodology, but they are two specific studies. They have both been published, so one can find much more complete descriptions of the work there.[3,4]

9-ANILINOACRIDINE ANTI-TUMOR COMPOUNDS

The first example I will cover is a set of 9-anilinoacridines, which are anti-tumor compounds. As you may have seen in the literature,[5] Denny, Cain, and co-workers have synthesized at least 1,000 9-anilinoacridines with a tremendous variety of substituents R_1 and R_2, as shown in Figure 3. Denny has been interested in looking for relationships among this very large data set, and he has spent a substantial amount of time learning about quantitative structure-activity relations (QSAR) and how to apply it to his problem.

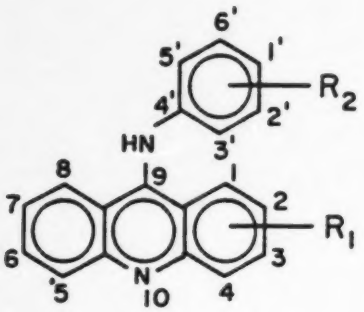

FIGURE 3. The 9-anilinoacridine's general structure.

In this data set, toxicity and potency are fairly highly related. It is possible to make more potent compounds in an anti-tumor assay, but they also have greater toxicity. What is sought is more selective compounds. Selectivity is reported for these data as the increase in life span at a fixed dose level. The fixed dose used in their studies has been LD_{10}, and so ILS_{max} is related to the tumor cell selectivity and is independent of toxicity. In the data set of more than 700 compounds provided to us by Denny, the selectivities ranged from none to very large values.

Attempts to generate quantitative models to predict ILS_{max} based on physicochemical parameters have not, to my knowledge, succeeded, and Denny suggested looking at this problem using the pattern recognition approach that I have been describing. To attack this problem using a discriminant, we put a threshold through the data at 35% ILS_{max}. That is, an active compound is one that has an ILS_{max} greater than 35% compared to controls, and an inactive compound is one that has less selectivity than that.

Denny supplied us with a list of 776 compounds from which a random selection was made. A set of 213 compounds comprised the data set to be used to try to develop discriminants. Within this set there are roughly twice as many active as inactive compounds—153 active compounds and 60 inactive compounds. Thus, the problem addressed was: Given 213 compounds with each tagged as good or bad, develop a set of descriptors that represent the compounds and will support a useful discriminant.

After a great deal of searching through descriptor sets, we settled on the set of 18 descriptors listed in Table 1. They range in complexity from the very simple—the number of sulfur atoms, the number of rings—to substructure environment descriptors and some partial charges at various positions in the molecule. These descriptors represent a diversity of properties in these molecules which, when taken as a group, can quite adequately describe a set of over 200 of these anti-tumor compounds and support a discriminant.

The substructures used to calculate the environment descriptors are shown in Figure 4. They cover the range from a fairly large substructure which is, in essence, the entire kernel of the molecule, through very small substructures such as the primary amino group. In each instance, what is being coded is the steric surroundings of this substructure as imbedded within the individual 213 structures of the data set.

This set of 18 descriptors, taken as a set, is sufficient to support a discriminant that can separate almost all of the actives from inactives. Random expectation would be 50% of the active and inactive class correct, and 50% correct overall. If one knew only that the data set was comprised of 153 actives and 60 inactives, he/she could use that information to guess, and one would get 72% in the active class and 28% in the inactive class correct, for an overall success rate of 60%. In fact, in the actual studies using the 18 descriptors, we got 94% overall with 97% correct in the active class and 85% correct in the inactive class. This means 200 correct classifications out of 213 compounds.

We then performed the usual internal consistency checks. Finally, we performed some actual predictions with unknowns. The results obtained for the unknowns were somewhat lower than for the original

TABLE 1
Descriptor Set for the 213 Compound 9-Anilinoacridine Data Set

No.	Descriptor[a]	Active Class Mean	Active Class SD	Active Class NNZ[b]	Inactive Class Mean	Inactive Class SD	Inactive Class NNZ	Highest R^2 Value Simple[c]	Highest R^2 Value Multiple[d]
1	No. of S atoms	0.804	0.415	122	0.417	0.530	24	0.745(12)	0.797
2	No. of rings	4.137	0.345	153	4.100	0.303	60	0.612(3)	0.898
3	Av no. of paths per atom[e]	107.9	20.30	153	94.14	18.98	60	0.707(4)	0.953
4	Molar refractivity	120.5	12.53	153	109.2	15.62	60	0.707(3)	0.925
5	Environment SS 1[f]	0.361	0.529	49	0.410	0.545	22	0.373(8)	0.596
6	Environment SS 2	0.452	0.824	36	0.246	0.636	8	0.112(7)	0.338
7	Environment SS 3	0.800	1.405	38	0.591	1.263	11	0.236(4)	0.663
8	Environment SS 4	0.358	0.745	29	0.188	0.569	6	0.373(5)	0.646
9	Environment SS 5	0.688	1.126	45	0.413	0.854	12	0.616(10)	0.706
10	Environment SS 6	0.482	0.934	33	0.276	0.772	7	0.616(9)	0.745
11	Environment SS 7	2.209	2.416	70	1.529	2.268	19	0.712(13)	0.857
12	Environment SS 8	3.351	2.206	113	1.567	2.164	21	0.745(1)	0.817
13	Environment SS 9	2.136	2.906	54	1.532	2.680	15	0.712(11)	0.870
14	Environment SS 10	4.097	3.416	91	6.221	1.942	55	0.138(12)	0.318
15	Environment SS 11	1.687	3.168	34	2.637	3.627	21	0.247(17)	0.572
16	Charge position 2[g]	−0.016	0.034	153	−0.015	0.039	60	0.187(8)	0.310
17	Charge position 3	0.028	0.059	153	−0.004	0.048	60	0.247(15)	0.449
18	Charge position 2′	−0.019	0.027	153	−0.008	0.044	60	0.021(15)	0.061

[a]Overall F statistic for between-group separation (based on Hotelling's T^2) is 7.03 (18, 194 df). [b]Number of nonzero values. [c]Highest R^2 with any other single descriptor (descriptor number in parentheses). [d]Multiple squared correlation with all other descriptors in table. [e]Total number of paths in molecule (of all lengths)/total number of atoms (nonhydrogen). [f]Substructures for environment descriptors in Chart I. [g]Partial σ electronic charge calculated by del-Re method.

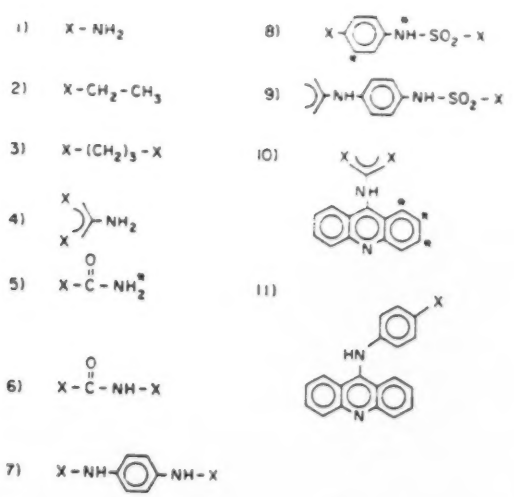

FIGURE 4. The substructures used for development of the descriptors for the 9-Anilinoacridine study. (An X indicates an unspecified atom type, and starred atoms are not checked for substituents during substructure searching.

data set, which is almost always the case in studies such as these. Obviously, predictions are much harder to make than retrospective classifications. In addition, the compounds forming the two prediction sets had some characteristics that should be pointed out.

The first prediction set contained 50 compounds taken out of the original 776 that Denny sent us. Therefore, it contained similar structures to those used in the data set for generation of the 18 descriptors and the discriminants. Of the 35 active compounds, 30 (86%) were correctly classified; of the 15 inactive compounds, 11 (73%) were correctly classified. This is 82% overall, which greatly exceeds the 58% that would be expected for guessing. And although this number is somewhat lower than that obtained for classification of the original data, it is still quite respectable.

For Prediction Set 2, the results were not as good. Sixty-nine compounds were taken out of two recently published papers in the *Journal of Medicinal Chemistry*.[6,7] They comprised all the compounds that appeared in the two papers on the subject of 9-anilinoacridine anti-tumor compounds that had the property of not being in our original data set. They are structurally somewhat different from the compounds that were in our original data set, and the results were degraded somewhat. Of the 45 actives, 35 (78%) were correct; of the 24 inactives, 18 (75%) were correct; that is, 77% overall. The degree to which these compounds were different is difficult to quantify because structural similarity is not a very quantitative concept. Investigations of structural similarity and how this relates to prediction success are in progress in the research group at this time.

N-NITROSO CARCINOGENS

My second example concerns N-nitroso compounds, which are an infamous class of chemical carcinogens. The N-nitroso compounds in this particular study[4] numbered 150 taken from the literature. Each compound was labeled as a carcinogen or benign compound. Sometimes there was additional information regarding potency, eg, strong or weak carcinogen. The data set contained 122 carcinogens and 38 non-carcinogens, covering a wide diversity of structural types. The only similarity among the compounds is that they all contain an N-nitroso group, by definition. In this particular study, a set of 22 descriptors was found which supported a discriminant of substantial separating power for carcinogens vs non-carcinogens. Their identities are shown in Figure 5. The descriptors ranged in complexity from something as simple as the number of nitrogen atoms through levels of complexity: molecular connectivity, some shape information, a lot of electronic information, molar refractivity, and a count of the number of hydrogens on the β carbons. We know from the literature that the metabolism of N-nitroso compounds often involves β hydrogens. We have a diverse set of descriptors, once again, representing these structures.

The substructures used for developing environment descriptors are shown in Figure 6. Four of them involve the N-nitroso group and are clearly coding either the electronic nature of the region of the molecule where the N-nitroso group is found or else the steric properties of this region. The other two substructures are coding branchings off benzene rings contained within the n-nitroso compounds' structures.

The classification results of this particular study were quite good. A number of different pattern recognition methods were applied to the same data set. This is one of the advantages of ADAPT. You can take the same data set, store it away, and apply a wide variety of pattern recognition methods to the same problem, searching for the ones that work best. The best results in this particular instance were 146 correctly classified out of 150 compounds – 97.3% correct classifications. So, this is a very large and somewhat diverse data set of N-nitroso compounds, essentially completely

Twenty-two Descriptors Giving 97% Classification of the 150 Compound Nitrosamine Data Set.

	NAME	DESCRIPTOR	MEAN	STD DEV	NNZ
1)	FRAG 4	Number of N atoms	2.31	0.685	150
2)	MOLC 4	Path 2 molecular connectivity	2.59	0.982	150
3)	GEOM 3	Intermed. principal moment	0.138	0.138	150
4)	GEOM 4	Largest PM/Intermed. PM	4.23	3.18	150
5)	ENVR 1	Mol. Conn. Envir. of SS 1	2.39	0.378	150
6)	SCAV 1	Average sigma charge of SS 1	-0.0490	0.00684	150
7)	SCMP 2	Most positive sigma charge of SS 2	0.00974	0.0232	25
8)	MREF 1	Molar Refractivity	41.2	12.4	150
9)	CSEP 1	Charge Separation	3.51	1.29	150
10)	SCMP 3	Most positive sigma charge of SS 3	0.0442	0.0223	121
11)	NOHC 2	Number of H on beta carbons	3.73	2.81	137
12)	SSS-2	Number of SS 2 occurrences	0.233	0.595	29
13)	PATH-2	Number of paths from SS 2	11.7	28.8	29
14)	SSS-3	Number of SS 3 occurrences	1.46	0.974	121
15)	PATH-3	Number of paths from SS 3	37.7	26.4	121
16)	SSS 2	Number of SS 2 occurrences (ring specific)	0.0933	0.335	12
17)	SSS 3	Number of SS 3 occurrences (ring specific)	0.787	0.909	72
18)	PATH 3	Number of paths from SS 3 (ring specific)	17.8	20.8	72
19)	SCMN 3	Most negative sigma charge of SS 3	-0.00506	0.0112	26
20)	SSS-4	Number of SS 4 occurrences	0.0733	0.262	11
21)	SCMP 5	Most positive sigma charge of SS 5	-0.00026	0.0166	26
22)	SCAV 6	Average sigma charge of SS 6	0.00668	0.0202	26

* NNZ = Number of Nonzero Occurrences

FIGURE 5. Descriptor set for the 150 compound N-nitroso compound data set.

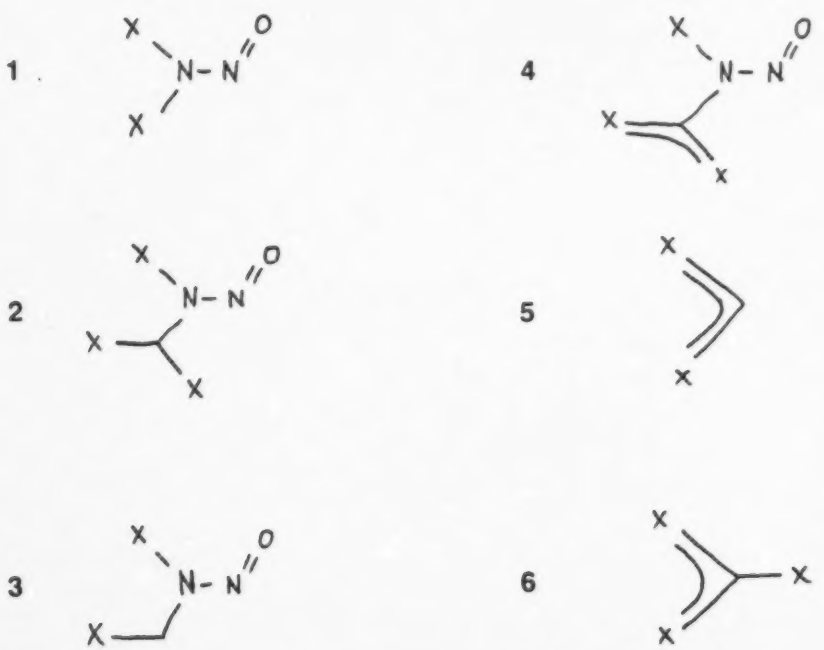

FIGURE 6. The substructures used for development of the descriptors for the N-nitroso compound study.

classified by this discriminant based on 22 calculated molecular structure descriptors.

SUMMARY

My objective here has been to introduce an approach to structure-activity studies that can be used to tackle problems with data sets that may be large, where the structures within those data sets may be fairly diverse, and where the biological activity data may be available only as category tags. The objective often is to generate discriminants rather than quantitative models. We have developed a software system that supports these kinds of studies and makes it easy for a researcher to step through a study of this type without having to keep a complex set of notes on paper; the software stores all the relevant information. So, chemical structure information handling and pattern recognition methods can be linked together to develop an approach that has some differences from other alternative approaches the SAR studies and should have application in a wide variety of different areas.

REFERENCES

1. Stuper AJ, Brugger WE, Jurs PC: *Computer Assisted Studies of Chemical Structure and Biological Function.* New York: Wiley-Interscience, 1979.
2. Hansch C, Leo A: *Substituent Constants for Correlation Analysis in Chemistry and Biology.* New York: Wiley-Interscience, 1979.
3. Henry DR, Jurs PC, Denny WA: Structure-antitumor activity relationships of 9-anilinoacridines using pattern recognition. *Jour Med Chem* 1982; 25:899–908.
4. Rose SL, Jurs PC: Computer-assisted studies of structure-activity relationships of N-nitroso compounds using pattern recognition. *Jour Med Chem* 1982;25:769–776.
5. Denny WA, Cain BF, Atwell GJ, Hansch C, Panthananickal A: Potential antitumor agents. 36. Quantitative relationships between experimental antitumor activity, toxicity, and structure for the general class of 9-anilinoacridine antitumor agents. *Jour Med Chem* 1982;25:276–315.
6. Baguley BC, Denny WA, Atwell GJ, Cain BF: Potential antitumor agents. 34. Quantitative relationships between DNA binding and molecular structure for 9-anilinoacridines. Substitution in the anilino ring. *Jour Med Chem* 1982;25:170–177.
7. Baguley BC, Denny WA, Atwell GJ, Cain BF: Potential antitumor agents. 35. Quantitative relationships between antitumor (L1210) potency and DNA binding for 4′-(9-acridinylamino)methanesulfon-m-anisside analogues.

NOTICE

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

This article documents research that was conducted under the auspices of the U.S. Environmental Protection Agency (USEPA). As an arm of the government, the USEPA has a paid up, non-exclusive, irrevocable, worldwide license to reproduce, prepare derivative works, and distribute to the public all of the material contained herein, and to grant others acting on its behalf the right to do so. The USEPA grants to the National Technical Information Service, Department of Commerce, the right to act as its agent in the distribution of this material.